

Explainable AI

Moving AI from Black Boxes
to Glass Ones

MUNSHI IMRAN HOSSAIN

Senior Data Scientist

at

Cytel Statistical Software & Services Pvt. Ltd.

Agenda

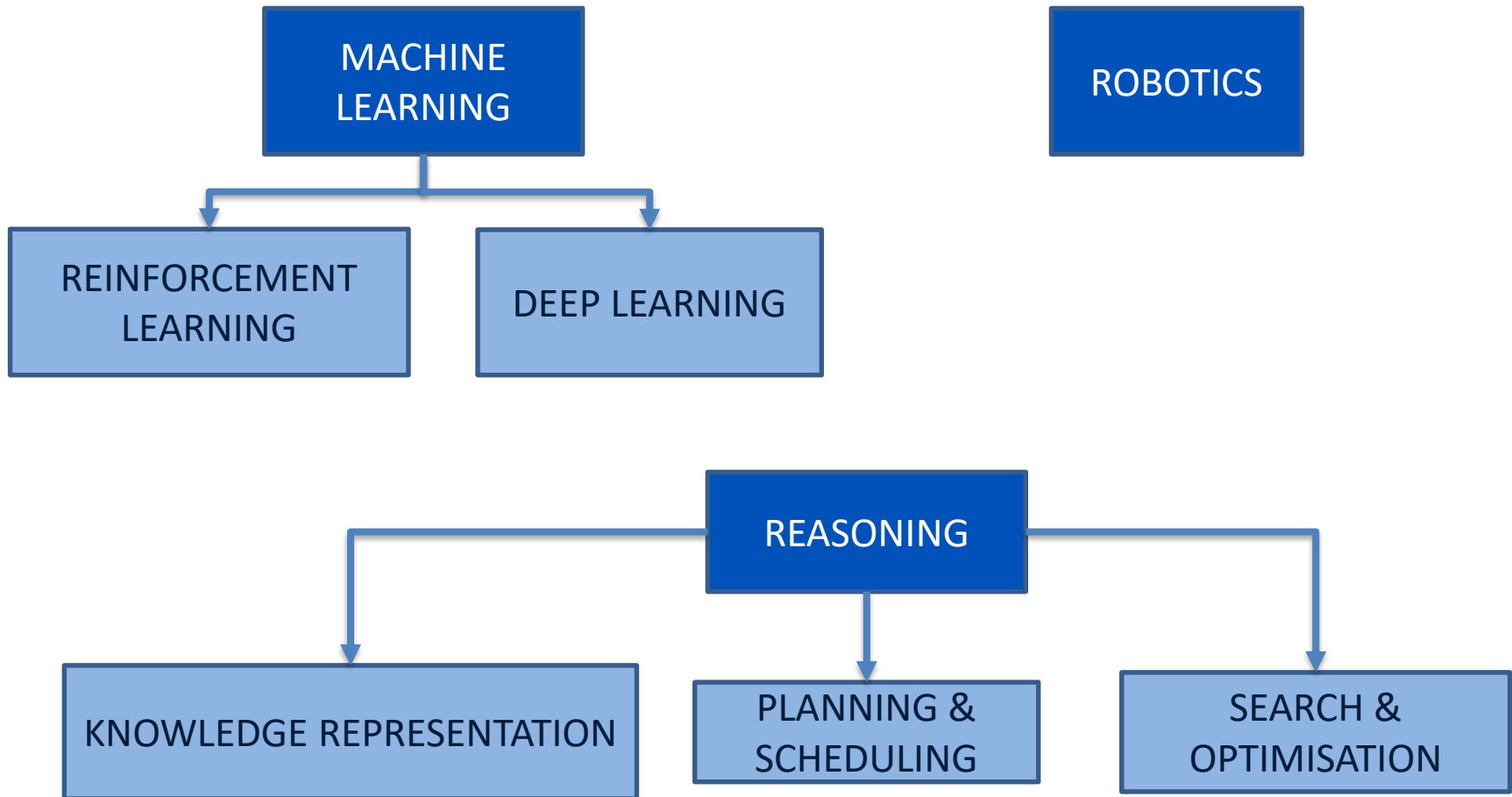
- What is AI?
- What is Black Box AI?
- Drawbacks of Black Box AI
- What is Explainable AI (XAI)?
- What Is The Need for XAI?
- Examples of XAI Methods

Disclaimer: *All opinions expressed in this presentation are that of the author and Cytel does not bear any responsibility.*

What is AI?

- AI is the acronym for Artificial Intelligence.
- *Artificial intelligence refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.*
 - The European Commission's high-level expert group on AI

The Sub-disciplines of AI



What is Black Box AI?

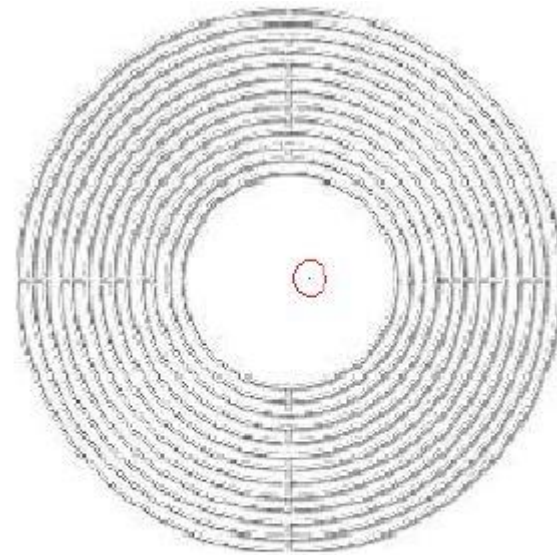
- Some AI techniques are very opaque in terms of understanding how they make decisions.
- Why did the system make a particular decision and why did it not do something else?
- When did the system succeed and when did it fail?
- How can the system correct an error?

Drawbacks of Black Box AI

- We don't really know why the system makes the choices it does.
- This lack of transparency is a problem for a clinician who wants to understand the reason behind the decision.
- GDPR introduced the right of explanation for individuals to obtain “*meaningful information of the logic involved*”. This will outlaw such systems.

One Pixel Attack on Deep NN

- Deep neural networks (NN) can be easily made to alter their decision by modifying only one pixel in the input!
- About 68% images in the CIFAR-10 test data and 41% images used in ImageNet test data, can be perturbed by at least one target class.



Whorl

Blower (37.00%)

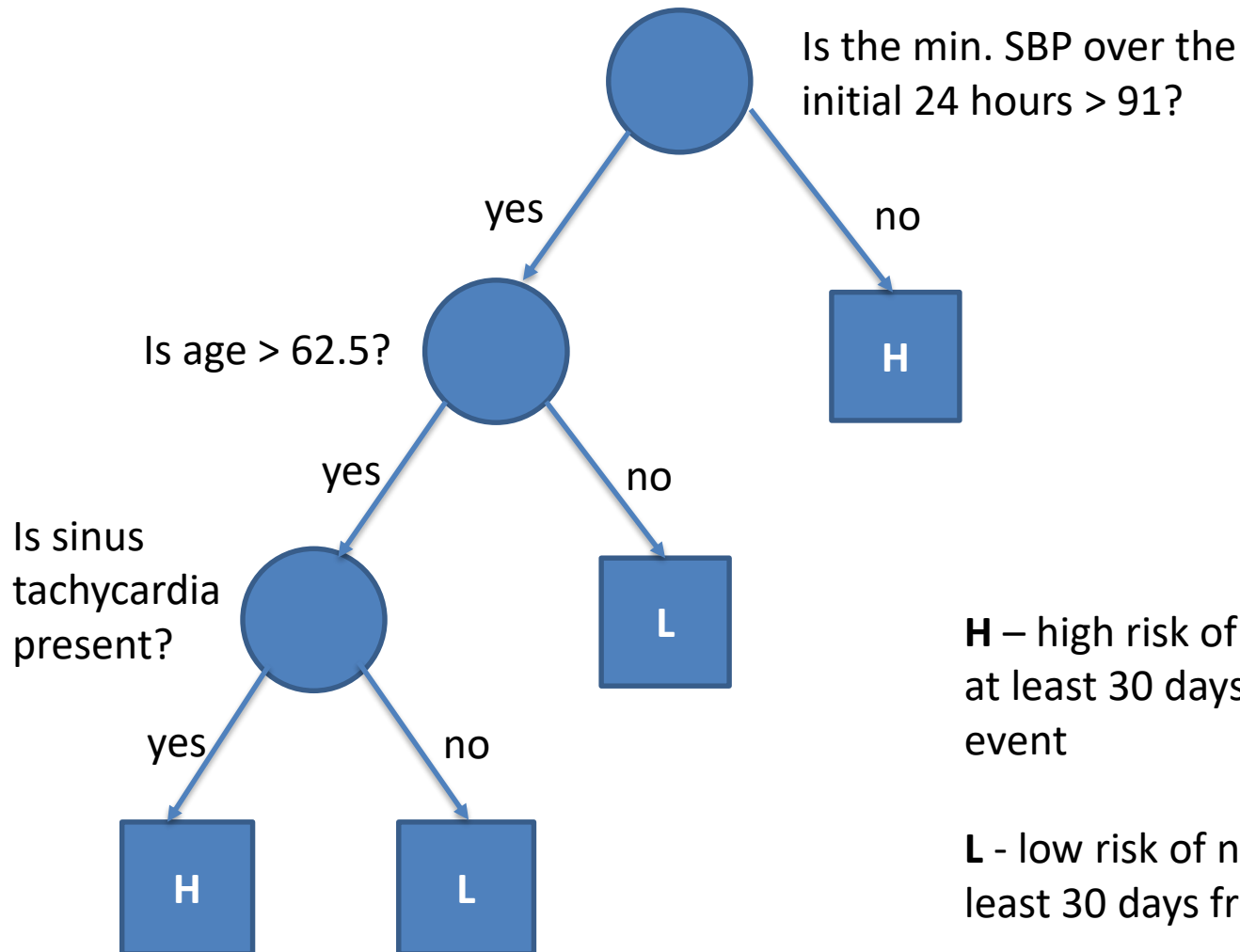
Algorithms With A Racial Bias

- In 2016, Amazon offered same-day delivery service to Prime users in major US cities.
- The scheme covered almost all zip codes, except the majority black ones.



What is Explainable AI (XAI)?

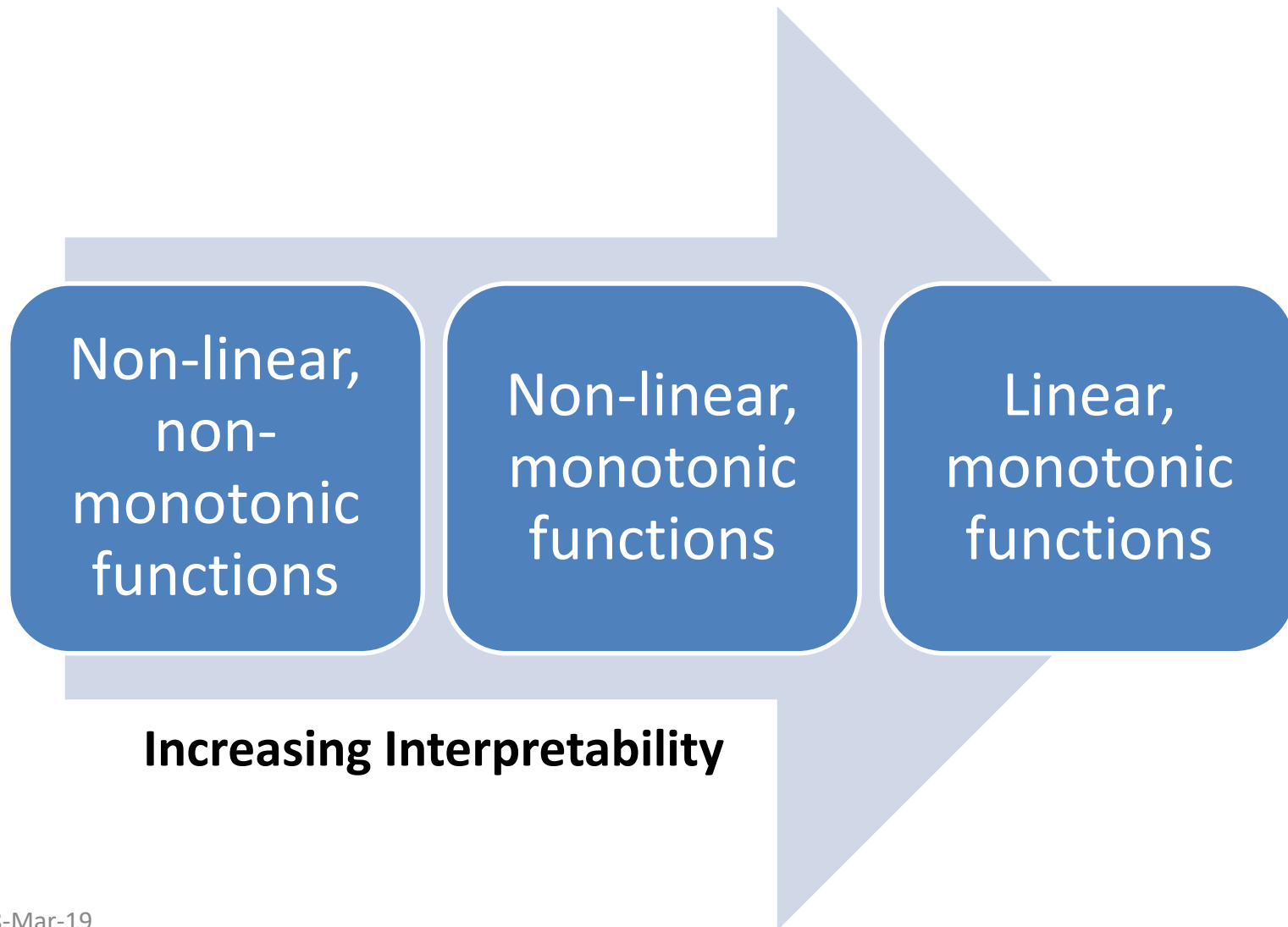
- The ability to explain a machine learning prediction
- Produce explainable models, while maintaining a high level of prediction accuracy
- Enable humans to understand, trust and manage the emerging generation of artificially intelligent systems



What Is The Need for XAI?

- Building trust to allow stakeholders to adopt these systems
- Preventing adversarial attacks on machine learning systems
- Being in line with industry regulations

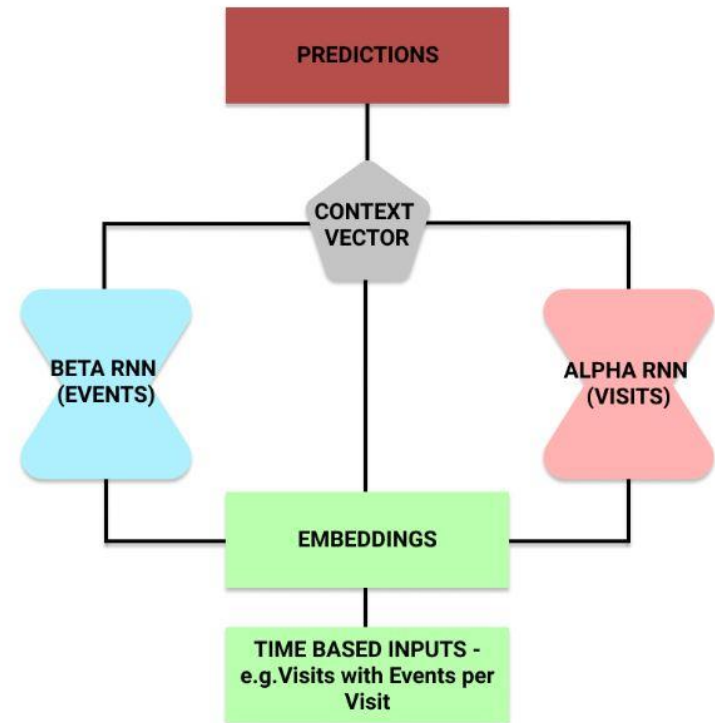
How Explainable Are Models?



Examples of XAI Methods

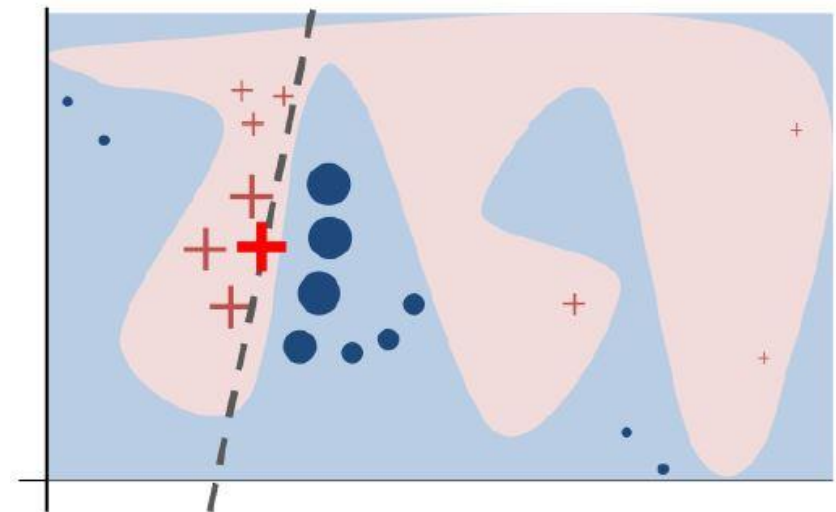
REverse Time AttentionIoN model (RETAIN)

- It is based on a 2-level neural attention model that detects influential past visits and the key diagnoses of those visits.
- The model relies on an attention mechanism modelled to represent the behaviour of physicians.



Local Interpretable Model-Agnostic Explanations (LIME)

- The objective is to identify an interpretable model that is locally faithful to the classifier.
- The idea is to perturb the inputs and watch how doing so affects the model's outputs.



Summary

- Black box AI models, though highly accurate, are not explainable.
- Today's cutting-edge data science algorithms need to be explainable before they are fit for adoption.
- This is important so that algorithms are in line with human values and expectations and preserve human autonomy and awareness in decision making.

References

- The European Commission's High-Level Expert Group On Artificial Intelligence, Brussels, 18 December, 2018
- Su, J., Vargas, D. V., Sakurai, K., 2019, One pixel attack for fooling deep neural networks
- Amazon Doesn't Consider the Race of Its Customer. Should It? - <https://www.bloomberg.com/graphics/2016-amazon-same-day/>
- Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., Sun, J., 2016, RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism, 29th Conference on Neural Information Processing Systems
- Ribeiro, M. T., Singh, S., Guestrin, C., "Why Should I Trust You?" Explaining the Predictions of Any Classifier, 2016, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144
- Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Pappalardo, L., Ruggieri, S., Turini, F., 2018, Open the Black Box: Data-Driven Explanation of Black Box Decision Systems, Journal of Computer Engineering & Information Technology, Vol. 1, No. 1

Thank you!

Contact: Munshiimran.Hossain@cytel.com